PROCEEDINGS OF TDWG

Biodiversity Information Science and Standarts 1: e20306
doi: 10.3897/tdwgproceedings.1.20306

OPEN ACCESS

Conference Abstract

Automated Generation of Lists of Unique Values from iDigBio Data Fields to Facilitate Data Quality Improvements

Saniya Sahdev‡, Deborah Paul§, Matthew Collins‡, Jose Fortes‡

‡ University of Florida, Gainesville, United States of America
§ Florida State University, Tallahassee, United States of America

Corresponding author: Saniya Sahdev (saniyasahdev@ufl.edu)

Received: 15 Aug 2017 | Published: 15 Aug 2017

Citation: Sahdev S, Paul D, Collins M, Fortes J (2017) Automated Generation of Lists of Unique Values from iDigBio Data Fields to Facilitate Data Quality Improvements. Proceedings of TDWG 1: e20306. https://doi.org/10.3897/tdwgproceedings.1.20306

Abstract

iDigBio currently has over 100 million records with up to 260 fields per record [Matsunaga et al. 2013]. Many of these fields are mapped to the Darwin Core (DwC) and Audubon Core standards. How well do the data in those fields meet the term definitions of those standards? Amassing biodiversity collections data into very large aggregated datasets offers never-before-possible ways in which to use the existing data to enhance current data and improve future data. While most data providers attempt to adhere to the recommended standards, looking inside the data entered for a given field across aggregated datasets has revealed significant data quality issues. Among other issues, data might be the wrong data type, mapped incorrectly, use old terminology, be formatted incorrectly, or use a non-standard controlled vocabulary.

The Darwin Core Hour webinar initiative [Zermoglio et al. 2017] started in January of 2017 to improve DwC implementation and documentation, as well as community engagement and understanding of the DwC standard and standards process. As part of that process, it was recognized that while aggregators with informatics skills can easily see the above data issues, it is not simple for most data providers or downstream users to visualize large

datasets. In fact, it is often difficult for data providers to visualize issues in their own local datasets.

One place to start improving data quality is with the fields from the DwC standard that recommend the use of a controlled vocabulary. There are 23 fields that recommend the use of a controlled vocabulary. A call went out to large aggregators to share comma separated values (CSV) files containing a list of distinct values found in each of these 23 fields, along with a count. The responses from iDigBio, the Global Biodiversity Information Facility (GBIF), and VertNet are stored in the TDWG Darwin Core Q&A GitHub repository [Paul 2017].

Based on this community need to have more insight into controlled vocabulary data as well as experience with iDigBio's existing data cleaning approaches, we have constructed an automated process to generate lists of unique values in iDigBio fields. We used the data available from dumps of the entire iDigBio data set, which are written out weekly and stored on the GUODA (Global Unified Open Data Access) infrastructure [Collins et al. 2017], the distributed processing engine Apache Spark, and the job management software Jenkins. The resulting CSV files are archived and automatically made publicly available once a week through the web on iDigBio's Ceph object store.

Dynamically generating this distinct value data is a first step in understanding the current vocabularies in use by data providers. Using summarization and clustering algorithms, data in the fields can be easily visualized and analyzed. With these data, not only can patterns beyond typos and counts be seen by anyone, but metrics can be put in place. As discipline-specific communities are able to easily see what is in a given field, they can work together to synthesize recommended vocabularies to improve future data. As the data are improved, the number of distinct clusters would be expected to decrease, as would the number of values found in a given cluster. Without these kinds of automated tools that build data products from aggregated data, it would be much harder to tackle many data quality issues.

Keywords

Biodiversity, Data Quality, Data Cleaning, Darwin Core, Cloud Computing, Bio Collections Infrastructure

Presenting author

Matthew Collins

References

- Collins M, Hammock J, Poelen J, Thessen A, Thompson A (2017) http://guoda.bio/about/. Accessed on: 2017-7-19.
- Matsunaga A, Thompson A, Figueiredo R, Germain-Aubrey C, Collins M, Beaman R, MacFadden B, Riccardi G, Soltis P, Page L, Fortes JB (2013) A Computational Storage Cloud for Integration of Biodiversity Collections. 2013 IEEE 9th International Conference on e-Science https://doi.org/10.1109/escience.2013.48
- Paul D (2017) https://github.com/tdwg/dwc-qa/blob/master/data/readme.md. Accessed on: 2017-7-19.
- Zermoglio P, Paul D, Krimmel E, Motz G, Wieczorek J (2017) https://www.idigbio.org/content/evolution-darwin-core-hour. Accessed on: 2017-7-19.